# Towards an International PhD in Computational Linguistics

A. Zampolli, L. Cignoni

## 1. Premise

The field of "Informatique Linguistique" (Computational Linguistics: CL) is so various that a generalized comprehensive definition of this discipline has not yet been given.

The wide variety of terms used to designate it in the different languages confirms the different interests and applications which characterize the researchers active in this field.

We feel it is instead useful to underline some particular aspects:

- The interdisciplinarity of CL, owing to the contribution given in many ways by different disciplines. These disciplines are - generally speaking - located within different didactic and academic frameworks with slight, if not totally lacking, 'administrative' intercommunication.
- The strong link between theoretical and applicational aspects.

CL deals in part with the application of linguistic theories to mechanized natural language processing systems directed towards specific "engineering" goals of the "industries de la langue", but its intellectual significance is much wider than this would suggest.

"It serves as a focus for theoretical work in many related disciplines, including Logic, Philosophy, Psychology, Theoretical Computer Science, and Speech and Communications. The field serves as a two-way channel of influence; from linguistic theory to computational applications, and from computational and cognitive theories of reasoning processes to models of language and language use".

## Multilinguism

The multilingual dimension is an important aspect of the language industry, especially in Europe and in the present framework of the so-called "information society".

## Research Activities and Training

The inclusion of the students within experienced research groups is the most efficient and, perhaps, an essential tool for training in computational linguistics at the PhD level.

Only the collaboration among universities of different countries - in which major research activities in computational linguistics are underway - can make available to the students a set of competences,

facilities, data and activities covering potentially the entire range of CL and the necessary multilingual dimension.

2. Possible Topics

The applicants to the PhD courses usually:

- come from different University backgrounds (Ital. "Facolta'": linguistics, computer science, humanity disciplines, philosophy, psychology, etc.)
- require specialization in a particular application field for which a specific and strongly differentiated training is essential: speech processing, machine translation, etc.

The following types of courses seem to be taken into consideration:

a) Prerequisites and Qualifying Courses

It seems essential that the students should start these courses having a minimal set of knowledges. After having assessed their level, the students should be offered the possibility of filling any possible gaps resulting from their University training, in the following sectors:

- linguistic theory
- aspects of computer science bearing on NLP
- logics

b) The core courses

The following topics seem to be natural candidates for core courses (1):

- Natural Language Processing
- Morphology
- Syntax
- Semantics
- Logic and natural Language
- Parsing
- Generation
- Lexicology and Lexicography
- Text Processing
- Pragmatics
- Cognitive Science
- AI and NLP; KR(L)

- Discourse analysis

- Hardware, software, Programming Languages for NLP

- Text processing and Literary Researches, Historical Linguistics , etc.

c) Areas of Specialization

For example

- Signal processing, Speech Synthesis and Recognition

- Contrastive Linguistics, MAHT, HAMT, MT

- Statistics, quantative linguistics, stylometrics

- Office automation

- Linguistic DB, IR, Documentation

- Dialogue, NL interfaces

- CL and aids for the handicapped

3. Possible contributions from Pisa

A. The following didactic activities are currently underway in Pisa:

A.1 UNDERGRADUATE LEVEL (Faculty of Letters)

Introductory course to computational linguistics

typical topics: literary and linguistic text processing; statistical linguistics; computational lexicology and lexicography; formal grammars; parsers and generators

Course in applied linguistics

typical topics: formal languages and automata, grammatical formalisms and parsing techniques

Computational linguistics tutorials

introduction to computer science for linguists; programming languages; DBMS; practical work on the CL data of the ILC.

A.2. POST-GRADUATE LEVEL

Curriculum in CL of the PhD in linguistics

The courses are organized as seminars and, mainly, as a theoretical/practical participation to the research activities of the computational linguistics groups operating both at the Department of Linguistics of the University and at the Institute of Computational Linguistics of CNR (see below).

Computational linguistics course held at the School of Specialization in "Computer Science".

Typical topics offered: natural language processing and knowledge representation

It should also be reminded that the Faculty of Computer Sciences of Pisa University was the first to be founded in Italy, and that most of the research activities of CNR in the sector of computer science (which accounts for approximately 300 researchers, distributed in three Institutes: CNUCE, IEI, ILC), are concentrated in Pisa.

B. Because we think that specialization at the PhD level should mainly take place through the participation to groups and advanced research activities, we shall now list the sectors for which Pisa is able to offer a suitable training, owing to the the national (Progetti Strategici) and international research activities (ESPRIT, EUREKA, EUROTRA, "LARPA", ecc.):

- literary and linguistic text processing
- computational lexicology and lexicography
- dialogue and natural language understanding
- linguistic DBs
- historical linguistics, classical languages and CL
- theory of grammars and parsers
- knowledge representation

C. The Institute of Computational Linguistics and Pisa University are well known for having organized a number of Summer Schools in computational linguistics, intensive interdisciplinary courses of 3/4 weeks, largely contributing to the formation of a large numer of researchers presently operating in Europe and, partly, in the United States.

The theme of the next Summer School, envisaged for 1988 and already sponsored by the European Science Foundation, will be Computational Lexicology and Lexicography.

D. Computational Facilities

IBM 3081/K; various specialized laser printers and plotters; SUN 3/185 with ETHERNET; various working stations (terminals, PCs, etc.); IBYCUS with CD-ROM.

E. Data and computational linguistics systems available

- Text Processing Procedures
- Computational syntax: grammatical formalisms and running parsers for natural language. Theoretical and computational models of dialogue
- Lisp, PROLOG for natural language processing
- Computational morphological analyzer (Italian, Spanish, Latin)
- Knowledge Representation Languages
- Italian Grammars
- Textual Data Base (50,000,000 words in 30 languages) and access procedures
- Machine Dictionaries and Lexical Data Bases (mono and bilingual)

(1) We shall not suggest a structure here (sequence, interdependence, specification of the contents), since we feel that it should emerge from the discussions which will take place during the meeting.

As an example, we are enclosing a list of the PhD courses held at Carnegie Mellon University:

Topics in Syntax
1. Generalized Phrase Structure Grammar
2. Government-Binding theory
3. Lexical-Function Grammar
4. Universal Grammar
5. Advanced Syntac Grammar

Topics in Semantics
1. Montague Grammar
2. Semantics of Programming languages
3. Advanced Semantics Seminar
4. Logical Semantics and Knowledge Representation

Topics in Logic
1. Inference and Natural Language
2. Automated Theorem Proving
3. Model Theory
4. Proof Theory

Topics in Speech Recognition
1. Acoustic Phonetics
2. Advanced Phonology
3. Speech perception
4. Principles of Automated Speech Recognition
5. Design of Speech Recognition Systems
6. Advanced Problems in Speech Recognition

Topics in Parsing
1. Parser Design and Parser Algorithms
2. Natural Language Processing and Computer-Aided Instruction
3. Intelligent Tutoring Systems
4. Grammar writing for Natural Language Processing
5. Problems of Machine Translation
6. Current Issues in Parsing

Topics in Lexicology
1. The Traditional Lexicon
2. Automated Electronic Dictionaries
3. Computational Morphology
4. Advanced Natural Language Morphology
5. Universal Natural Language Processing
6. Applications of Data Base Management and Knowledge Representation to the Lexicon

Topics in Pragmatics
1. Discourse Representation Theory
2. Philosophy of Language
3. Topics in Computational Processing of Pragmatic Phenomena
4. Advanced Seminar in Pragmatics

Topics in Cognitive Science
1. Learning Theory and Language Acquisition
2. Parallel Models of Cognition

Furthermore, we are reporting the following information extracted from the Survey promoted by the ACL. A copy of the drafts (kindly provided by D. Walker for our meeting) will be available at the meeting.

Location of the PhD academic courses reported in the Survey:
57 Computer Sciences
27 Linguistics
6 Psychology
5 Modern languages, Cognitive Sciences
4 Engineering

3 Computational Linguistics, AI
1 Philosophy, Behavioural Science, Library Science, Information Technology

We shall now report a list of the topics which are more frequently quoted in the above Survey.
Historic overview of CL
Relations of CL to : - Art.Intell.
- Philosophy
- Theoretical Linguistics
- Psychology

Parsing
Parsing strategies
Ambiguity and parsing
Deterministic and not deterministic parsing
TOP DOWN and BOTTOM UP parsing
Morphological analysis
Syntax analysis
Semantic analysis
Discourse analysis
Pragmatics
Language generation
Language comprehension
Syntax and semantic theories
Models of linguistic description
Formal language theories
Automata
Transformational grammar
Case grammar
Generative Semantics
Lexical Functional grammar
Montague grammar
Systemic grammar
Logic Grammars
Compositional semantics
Conceptual Dependency
Procedural semantics
NL representation
NL understanding
Text processing
Focusing
Anaphora
Semantic and contextual interpretation
Transition Networks and ATN
Semantic Networks
World knowledge
Knowledge representation languages
Theoreme proving
Problem solving approach
Logic for NL
Set theory
Formal logic
Cognitive Science
Cognitive aspects of language acquisition
NL in Art.Intell.
Lexicon
Lexicon-driven analysis
Dictionary construction
Lexicography
Thesaurus construction
Use of linguistic data bases
Machine translation
NL Question answering
Psycholinguistic
Sociological applications

Modelling of child language acquisition
Speech act theory
Philosophy of language
Statistical linguistic
Mathematical topics in NL processing
Concordances
Phonetics applications
Speech synthesis
Historical-comparative linguistics by computer
Software for NL processing
LISP
PROLOG
Application of CL